

文章编号: 1003-0077(2019)07-0110-08

基于语言学扰动的事件检测数据增强方法

陆垚杰^{1,2}, 林鸿宇^{1,2}, 韩先培¹, 孙乐¹

(1. 中国科学院 软件研究所, 北京 100190;

2. 中国科学院大学, 北京 100049)

摘要: 近年来,深度学习在事件检测领域取得了长足进展。但是,现有方法通常受制于事件检测标注数据的规模和训练阶段的不稳定性。针对上述问题,本文提出了基于语言学扰动的事件检测数据增强方法,从语法和语义两个角度生成伪数据来提升事件检测的性能。为了有效的利用生成的伪数据,该文探索了数据增加和多实例学习两个训练策略。在 KBP 2017 事件检测数据集上的实验验证了我们方法的有效性。此外,在人工构造的少量 ACE2005 数据集上的实验结果证明该文方法可以大幅度提升小数据情况下的模型学习性能。

关键词: 事件检测;数据增强;多实例学习

中图分类号: TP391

文献标识码: A

Linguistic Perturbation Based Data Augmentation for Event Detection

LU Yaojie^{1,2}, LIN Hongyu^{1,2}, HAN Xianpei¹, SUN Le¹

(1. The Institute of Software, Chinese Academy of Sciences, Beijing 100190, China;

2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Deep learning recently applied in the event detection task is limited by the scarcity of the annotated data and the instability during the training phase. This paper proposes a data augmentation method based on linguistic perturbation for event detection, which generates pseudo data from both syntactic and semantic perspectives to improve the performance of event detection systems. In order to effectively exploit generated pseudo data, this paper explores two training strategies: data addition and multi-instance learning. Experiments on the KBP 2017 event detection dataset demonstrate the effectiveness of our approach. Furthermore, the empirical results on a manual constructed portion of ACE2005 dataset show that the proposed method can significantly improve the model performance on small training data.

Keywords: event detection; data augmentation; multi instance learning

0 引言

事件抽取(event extraction)是自然语言处理的重要任务。事件抽取通常被划分为事件检测(event detection)和论元识别(argument recognition)两个子任务,其中,事件检测是事件抽取的基础。具体而言,事件检测的目标是检测出文本中最能代表事件类型的触发词。例如,在句子“An American tank *fired* on the Pales-tine Hotel.”中 *fired* 是句中表达的攻击事件的触发词。

近年来,深度学习模型在事件检测任务中取得了长足进展^[1-2]。相比于传统方法^[3],该类方法的优势在于避免了繁重的特征工程,同时其端到端的学习也避免了自然语言处理工具带来的错误传播问题。

深度学习模型要取得良好性能需要有大量的训练数据支撑。由于事件结构的复杂性,事件标注对标注者的专业要求高,且事件检测数据标注成本高,代价大。上述现象导致了事件检测的标注语料瓶颈问题。近年来,一部分研究者开始研究如何自动增加事件检测的数据,但这类方法通常是利用现有的

收稿日期: 2018-00-00 定稿日期: 2018-00-00

基金项目: 国家自然科学基金(61433015, 61572477, 61772505);中国科协青年人才托举工程(YESS20160177)

知识库(FrameNet、FreeBase)来实现^[4-5]。

另一方面,深度学习模型的训练过程也受到随机初始化等因素的影响。事件检测数据稀少,同时语言的表达具有稀疏性,进一步加重了上述模型训练过程中的不稳定性。为了提升模型训练过程的稳定性,研究者开始采用对数据或隐层表示注入噪声等训练方法^[6]。与此同时,受图像领域对抗实例研究的启发^[7],部分研究者也将对抗实例的方法应用于自然语言处理^[8],但在事件检测任务中尚未进行尝试。

为解决标注语料瓶颈问题和模型训练稳定性问题,本文提出了基于语言学扰动的伪数据生成方法,对事件检测进行数据增强。例如,我们可以通过词汇替换的方法,将句子“The nurse bandaged up his injured finger.”中的“injured”替换成“wounded”后的新句子“The nurse bandaged up his wounded finger.”加入到训练数据中进行数据增强。相比图像数据,自然语言的数据增强具有以下几方面的挑战。首先,自然语言表达是离散的、多样化的,简单地使用图像数据增强的方法会导致文本语义漂移的现象。其次,语言表达具有语法结构顺序,随意地替换文本片段会使文本不完整或意义发生改变。为解决上述挑战,在保留语义的情况下生成尽量多符合语法结构的伪数据是本文要解决的主要问题。具体地,本文借鉴之前的鲁棒训练相关工作^[9],从语法和语义的角度出发,提出了两种面向事件检测伪数据生成的语言学扰动方法,分别是基于语法的文本改写和基于语义的词汇替换。其中,基于语法的文本改写采用对文本进行复述或压缩的方式来生成伪数据,基于语义的词汇替换采用外部词汇语义资源来对原始文本词汇替换的方式生成伪数据。

为了验证伪数据的有效性,本文在 KBP2017^①事件检测数据集和人工构造的小规模数据集上进行了实验。实验表明,基于语言学的伪数据可以同时提升模型性能和训练过程的鲁棒性。

1 基于语言学扰动的事件检测数据生成

文本由离散的符号组成,其表达往往具有稀疏性。同一语义表达往往具有多种不同的文本表达形式。为了捕捉文本表达的多样性,提升有限训练数据下事件检测系统的性能,本文提出了基于语言学扰动的伪数据生成方法。具体地,本节介绍如何在保留文本语义的同时生成符合语法的事件检测实

例。借鉴 Li 等人的思路^[9],我们主要采用两类思路:基于句子改写的实例生成和基于词汇替换的实例生成。

1.1 基于句子改写的实例生成

句子改写的目的是,在可以在保留语义的情况下生成新的文本表达。例如,对于句子“*In Baghdad, a cameraman died when an American tank fired on the Palestine hotel.*”我们可以对其进行复述改写,得到新的句子“*An American tank fired on the Palestine led to a cameraman died in Baghdad.*”通过句子改写的方式,可以获得相同语义的多种文本表达。本文采用两种不同的句子改写方法对事件检测实例进行改写。

本文采用的第一种句子改写方式是基于文法的复述改写,利用 ERG (english resource grammar)^[10]实现。复述是对一段文本语义利用不同句法、词汇的重新表达。ERG 算法通过最小递归语义 (minimal recursion semantics, MRS)^[11]得到的语义表达式来对文本进行解析和生成。我们首先通过答案限制引擎 (the answer constraint engine) 解析器^②来对文本进行解析,然后通过预训练的解析结构选择模型来得到合适的解析结果,最后通过 MRS 来生成最后的合成结果。与原有数据相比,ERG 复述模型产生的新数据主要包含了形容词顺序、副词短语的位置变化。为了保留事件实例的语义,若改写后的文本不存在原有的事件触发词,则将该次改写视作一次不成功的改写,不加入到伪数据中。

本文采用的第二种句子改写方式是基于压缩 (Comp) 的句子改写。其出发点是在尽可能保留原有语义的情况下,生成长度较短、表达更加简洁的句子表达。例如,对于句子“*In Baghdad, a cameraman died when an American tank fired on the Palestine Hotel.*”中的攻击事件,其事件的语义集中在“sth. fired”。我们可以将其改写成“*A cameraman died when an tank fired.*”。可见,句子压缩可以在保证尽可能符合语法并保留原始语法结构,通过“剔除”原始文本中次要内容的方法来实现。由于缺少大量的训练数据,同时为了保证句子改写的精度,本文采用了一个基于语法解析器的句子压

① <https://tac.nist.gov/2017/KBP/data.html>

② <http://sweaglesw.org/linguistics/ace>

缩方法来对原始句子进行压缩。首先,我们对原始文本进行句法树分析。接着我们对从父节点标签为 R 的子树中随机删去具有标签 S 的子树 C 的概率进行了如下建模,如式(1)所示。

$$p(C | S, R) = \text{weight}_{comp} \frac{p(C, S, R)}{\sum_C p(C, S, R)} \quad (1)$$

概率模型 $p(C, S, R)$ 采用 Li 等的方法^[9],在 Clarke 和 Lapata 公开的句子压缩数据集^①上训练得到。为了尽可能保留原始文本中尽可能多的事件语义,我们通过 weight_{comp} 来控制事件语义的流失,若子树中含有真实的触发词,我们则将权重置为 0,反之则置为 1,如式(2)所示。

$$\text{weight}_{comp} = \begin{cases} 0 & \text{trigger} \in C \\ 1 & \text{otherwise} \end{cases} \quad (2)$$

1.2 基于词汇替换的实例生成

基于句子改写的实例生成可以有效生成具有不同语法结构的新实例。本文使用的第二种方法是基于词汇替换的实例生成。与基于句子改写的方法不同,词汇替换方法的核心是对用词层面的多样性进行建模。具体地,词汇替换方法需要解决两方面的问题:(1)如何选择被替换的词,才能保留尽可能多的语义?(2)如何选择新的词来替换原有词,不会使语义产生太大的漂移?

针对第一个问题,我们希望在替换词的过程中,尽可能保留事件实例原有的文本结构和与事件联系紧密的语义信息。功能词是决定句子结构的重要组成部分^[12],同时实体信息是决定事件类别的有效特征^[13]。因此本文保留功能词和实体标记范围内的词汇,只对其余词性标注为名词、动词及形容词等进行词汇替换。

针对第二个问题,我们希望用于替换的新词是与原有词在语义上接近,同时与局部上下文一致。因此,我们利用词汇语义资源来构建替换的候选词集 w_{cand} 。为选择与局部上下文相匹配的新词,我们采用统计语言模型来对替换后的文本片段进行语言模型打分,并通过打分 score_{lm} 来对候选词集中的词进行采样,生成伪数据。

上述过程的具体算法描述如算法 1 所示,为了选取与原词语义相近的新词,在算法 1 步骤 6 中,本文选取了两种不同词汇语义资源 Lexi 来生成新词候选,分别是 WordNet (WN)^② 和词嵌入表示 (CFit)。针对原词 w ,基于 WordNet 的候选生成方法将 w 的所有同义词作为候选集。词嵌入方法使

用词嵌入的相似度来选择候选。为了加强词嵌入中同义词之间的关系,本文采用 Count-Fitting^[14]的方法训练得到的词向量。该方法在原有 Word2Vec 模型^[15]的基础上,对反义词相似度添加惩罚,对同义词相似度进行增强,使得同义词的词嵌入相似度更高。

算法 1 基于语义的词汇替换

算法输入: 句子 sent , 语言模型 LM , 替换概率 α , 词汇语义资源 $Lexi$

1. $\text{sent}_{new} \leftarrow []$
2. For w in sent :
3. If w is function word or w is part of entity or w is Trigger
4. then $\text{sent}_{new} = \text{sent}_{new} + w$
5. $w_{cand} \leftarrow \emptyset$
6. For w_{new} in $Lexi(w)$:
7. $w_{cand} = w_{cand} \cup \{w\}$
8. $p(w) = \alpha$;
9. For w_{new} in w_{cand} :
10. $p(w_{new}) = (1 - \alpha) \cdot \frac{LM(w_{new}, \text{sent})}{\sum_{w_c \in w_{cand}} LM(w_c, \text{sent})}$
11. $\text{sent}_{new} = \text{sent}_{new} + \text{sample}(p)$

算法输出: 新生成的句子 sent_{new}

为了使替换的新词与原文的上下文更加匹配,在算法 1 步骤 8 中,我们利用候选词集中词与上下文的语言模型打分作为上下文的匹配程度,并通过随机采样的方法来选择新词。本文使用 KenLM 工具^[16]在 GigaWord^③上的 3 元语言模型作为实验所用的语言模型。其中, α 是替换词的阈值,当 α 越高时,越倾向于保留原有的词;反之,越倾向于替换成新的词。本文使用的 α 为 0.5。

2 基于增强数据的事件检测模型

本节首先介绍使用的基础事件检测模型,然后描述如何在训练过程中使用生成的数据对训练过程进行增强。

2.1 基础模型: 动态多池化卷积神经网络

与前人工作一致,我们将事件检测建模成一个多分类问题。给定一个句子,我们依次对句子中的每个词进行预测,判定其是否为一个事件触

① <http://jamesclarke.net/research/resources>

② <https://wordnet.princeton.edu>

③ <https://catalog.ldc.upenn.edu/ldc2009t13>

发词,同时将事件触发词划分到预先给定的类别上。本文使用动态多池化卷积神经网络(DMC-NN)^[1]作为实验的基础模型。如图1所示,该模型

能自动学习句子级别特征和词汇级别特征,主要包含三个模块:嵌入表示层、特征提取层和多层感知机分类器。

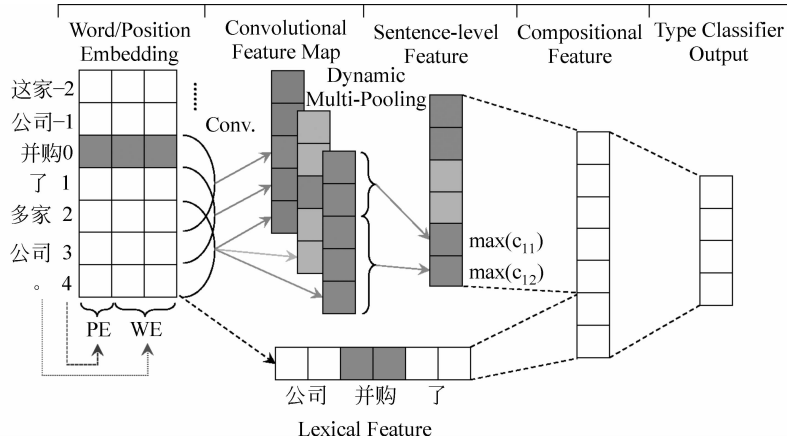


图1 动态多池化卷积神经网络结构

嵌入表示层对文本进行基础特征表示。本文主要采用两类嵌入特征来对文本进行表示,分别是词嵌入(word embedding, WE)^[15]和位置嵌入(position embedding, PE)。其中词嵌入主要捕捉词汇的语言学和语义特征,位置嵌入通过将上下文词与候选触发词的相对距离作为特征进行嵌入表示来建模不同上下文与候选触发词的关系。本文采用Skip-gram模型^①在Gigaword数据上对词嵌入进行预训练,同时对位置嵌入进行随机初始化,在训练过程中对两个嵌入表示同时进行更新。

特征提取层对文本表示完成事件特征提取。事件特征分为句子级特征和词汇级特征两部分。句子级特征由动态池化卷积层提取得到。动态池化卷积层首先对文本嵌入表示进行卷积操作获得卷积特征(convolutional feature map),并基于候选触发词位置对卷积特征进行动态分段多池化(dynamic multi-pooling),获得句子级特征(sentence level feature)表示S,用来表示候选触发词所在句子的全局信息。词汇级特征L(lexical feature)则将候选触发词及其特定窗口大小内的上下文词嵌入进行拼接,用来表示候选触发词的局部信息。

多层感知机分类器对经特征提取层得到的不同级别事件特征进行分类。我们将句子级特征和词汇级特征进行拼接,作为合成特征(compositional feature),共同输入到多层感知机分类器中,如式(3)所示。

$$O = \text{softmax}(W^{ds}[S;L] + b^{ds}) \quad (3)$$

其中, W^{ds} 是分类器权重矩阵, b^{ds} 是偏置项。输

出向量O代表不同类别的预测概率(Type Classifier Output),对于候选触发词t是类别j的概率如式(4)所示。

$$P(j | t, \theta) = O_j \quad (4)$$

其中, θ 是所有的模型参数。对原始的M个训练实例对 $(y^{(i)}, x^{(i)})$,我们采用最大化对数似然函数对基础模型进行训练,如式(5)所示。

$$L_{\text{base}} = \frac{1}{M} \sum_{i=1}^M \log P(y^{(i)} | x^{(i)}, \theta) \quad (5)$$

2.2 融入伪数据的模型训练

为了有效利用生成的伪数据来提升事件检测模型性能,本文使用两种不同的训练策略来对事件检测模型进行训练。

第一种方法将生成的N个伪数据看成真实的标注数据,直接将伪数据加入到原始的训练数据中,训练时最大化的目标函数如式(6)所示。

$$L_{\text{augu}} = \frac{1}{M+N} \left[\sum_{i=1}^M \log P(y^{(i)} | x^{(i)}, \theta) + \sum_{k=1}^N \log P(y^{(k)} | x^{(k)}, \theta) \right] \quad (6)$$

第二种方法借鉴多实例学习(multi instance learning, MIL)的思路^[4,17],采用多实例联合学习的方式来更有效的利用生成的训练数据和规避数据噪声带来的影响。具体地,我们将生成的伪数据依据事件类别划分成T个包 $\{M_1, M_2, \dots, M_T\}$,其中每

① <https://code.google.com/archive/p/word2vec>

个包中包含 q_i 个训练实例 $M_i = \{m_i^1, m_i^2, \dots, m_i^{q_i}\}$ 。多实例的学习任务是要对整个包的标签进行预测,因此我们在包级别定义了训练目标函数,如式(7)所示。

$$l_{ML} = \sum_{i=1}^T \log p(y_i | m_i^j, \theta) \quad (7)$$

其中, $p(y_i | m_i^j, \theta)$ 是针对第 i 个包中第 j 个实例利用参考式(4)的预测结果,我们对每个包中的实例进行事件类别概率预测,选取对真实标签预测结果置信度最高的实例 j 作为该包的预测结果,如式(8)所示。

$$j^* = \arg \max_j p(y_i | m_i^j, \theta) \quad (8)$$

最终,我们将基于真实数据的目标函数与基于伪数据的多实例目标函数进行联合训练,来对模型参数 θ 进行学习,如式(9)所示。

$$L_{ML} = L_{base} + \lambda \cdot l_{ML} \quad (9)$$

其中, λ 是平衡前后两项的权重因子,针对不同的伪数据,我们利用开发集从 $\{0.25, 0.5, 0.75, 1.0\}$ 中选择最优的 λ 。

3 实验及分析

为了验证本文方法的有效性,本文在 KBP2017 事件检测数据集和人工构造的极小数据集上分别进行了实验。

3.1 实验设置

本文使用 TAC KBP 2017 评测^①的事件检测数据集(LDC2017E55)作为测试集,共包含 167 个包含实体、关系和事件(Rich ERE)标记的英文文档。我们采用包含往年的 RichERE 标记数据集作为训练集,包括 LDC2014E31、LDC2015E29、LDC2015E68、LDC2016E31 和包含 TAC KBP 2015-2016 评测数据集的 LDC2017E02 作为训练集。为了进行超参数和模型的选择,我们从训练集中的 KBP 2016 评测数据中随机抽取 20 个文档作为开发集。最终,训练集/开发集/测试集划分数量为 866/20/167。我们采用斯坦福大学(Stanford)的 CoreNLP 工具^②对文档进行预处理,包括 XML 标签去除、句子划分和词条化(tokenize)、词性标注和句法树解析。

在嵌入表示层,本文采用的词嵌入维度是 300 维,位置嵌入维度为 5 维。在特征提取层,卷积神经网络的卷积窗口为 3,隐层大小为 300 维,词汇特征的窗口为 1(即使用候选触发词和左右词)。为了对

模型参数 θ 进行优化,我们采用 Mini-Batch 的方式来最大化训练目标函数,并采用 Adadelta 训练算法^[18]和 Dropout 的方式^[19]使得训练过程更加鲁棒,Dropout 概率为 50%。

为了验证基于语言学扰动的数据增强方法有效性,我们构建了基于随机词替换(WordDrop)的基准系统。WordDrop 受图像中的数据增强方式启发,随机选择被替换的词替换成 UNK。这样产生的伪数据噪声较大,出现很多不符合语法的、语义漂移的句子。我们依据模型在开发集上的性能选择了最优的 WordDrop 概率,本文使用 5%。

本文使用 KBP 2017 官方评测工具^③计算模型在测试集上的 F1 得分,并汇报每个模型运行十次所得事件类别预测的 F1 得分平均值。为了验证模型训练的鲁棒性,同时给出了十次实验之上的标准差。因为 TAC KBP2017 允许每一个队伍提交 3 个不同的运行结果。参考这一思路,我们依据开发集结果选择了 3 个最好的运行结果并取平均值,作为 Best-3 指标。

3.2 总体结论

不同数据生成方式和不同训练策略在 KBP 2017 数据集上的实验结果如表 1 所示。其中:DM-CNN 是采用目标函数 L_{base} 训练的模型,是本文方法的基准实验;ERG 和 Comp 分别是基于语法的句子改写方法中的基于 ERG 的复述方法和句子压缩方法;WN 和 CFIT 分别是基于 WordNet 和 CFit 词嵌入的词汇替换方法。其中,* 代表采用 MIL 方式进行的联合训练结果,+ 代表直接生成的伪数据加入到原始数据中的 Augu 方式进行训练。

表 1 KBP 2017 事件检测数据集实验结果

系统	平均值	标准差	Best-3
DMCNN (Baseline)	47.38	0.90	48.37
+ERG	47.63	0.62	48.33
+Comp	46.73	0.88	47.84
+WN	47.45	0.39	47.89
+CFIT	48.01	1.02	49.15
* ERG	47.14	0.57	47.94
* Comp	47.73	0.43	48.38

① <https://tac.nist.gov/2017/KBP>

② <https://stanfordnlp.github.io/CoreNLP>

③ <https://github.com/hunterhector/EvmEval>

续表

系统	平均值	标准差	Best-3
* WN	47.90	0.28	48.06
* CFIT	48.16	0.44	48.67
WordDrop	47.24	1.14	48.71

注：* 代表采用 MIL 方式进行的联合训练结果，+ 代表直接生成的伪数据加入到原始数据中的 Augu 方式进行训练。ERG\Comp\WN\CFIT 最优的 λ 分别为 0.5\0.5\0.75\0.25。

对比基准系统和加入伪数据后的平均值，8 个加入伪数据的结果中有 6 个系统(+ERG、+CFIT、+WN、* Comp、* WN、* CFIT)取得比基准系统更好的结果；有 1 个系统(* ERG)取得与基准系统相近的结果，但标准差更小。这表明我们生成的伪数据可以达到提升模型性能的作用。

对比实验结果的标准差指标，我们可以发现语言学扰动使得训练更加鲁棒。相比于基准系统的标准差 0.90，加入伪数据的方法普遍取得了更低的标准差。对比不同训练策略，我们发现使用基于 MIL 策略的伪数据训练方式能获得更小的标准差。这说明使用 MIL 策略可以有效地降低加入生成的伪数据进行训练所带来的噪声。与这两种方法相比，WordDrop 也可以达到改变文本，增加文本多样化的作用。但这样产生的文本更加不可控，语义漂移也较大。因此，标准差较大，使模型训练鲁棒性下降。

本文提出的方法并没有引入外部数据来进行训练，只是通过语义扰动的方式，对原始数据周围的实体特征空间进行了探索，生成了伪数据。

3.3 对比分析

针对不同语言学扰动方法的对比：基于词汇替换的扰动方法在所有实验中取得了最好的结果(包括均值、标准差和 Best-3)。因为 KBP 数据是新闻数据和论文数据的混合，而论文文本大多不太规范。基于句子改写的扰动方法在对句子改写的过程中，需要用到原始文本的句法解析树的结果，不规范的文本表达会导致句法树解析错误的增加。这样的错误会传播到生成的伪数据中，导致生成的数据丢失原有语义或者生成不符合语法的伪数据，为训练过程带来更大噪声。基于语义的词汇替换虽然取得了较好的效果，但这类方法仅仅是对词汇进行了替换，难以应对更多样化的文本表达。

针对不同伪数据训练策略的对比：采用基于 MIL 的联合目标训练策略在大部分实验中都取得

了更小的标准差，使得模型训练更加鲁棒。这说明基于 MIL 的训练策略可以更好应对伪数据中潜在的噪声。与之相比，基于 Augu 方式的训练策略，将每一个生成的伪数据都看成真实标注数据，忽略了伪数据本身有可能存在的不正确性。从而使得训练存在一定程度的不稳定性，但由于随机性增大，也会使得模型有可能在随机优化的过程中，取得更好的结果。基于 MIL 的训练策略通过包的误差来代表总体的误差，这样可以使模型自动规避置信度较低的实例，达到鲁棒训练的目的。

3.4 小规模数据实验

由于 KBP 训练数据集本身标注规模较大，数据增强的提升空间不大。为了验证本文提出的方法在提升小规模训练数据上的效果，我们人工构建了一个基于 ACE2005 数据^①的数据集以进行实验。相比于 KBP 数据，ACE 数据规模更小，同时我们进一步将 ACE 数据进行缩小，只保留一半的训练数据进行训练，同时在完整的开发集和测试集数据划分^[1]上进行训练。依据章节 3.3 的结论，我们采用基于语义的词汇替换方法来进行语言学扰动生成伪数据，同时采用多实例的训练策略进行训练。由表 2 中的实验结果表明，生成的伪数据可以对原来模型带来提升。

表 2 人工构建 ACE 数据实验结果

系统	平均值	Best-3
DMCNN (Baseline)	65.59	65.85
* CFIT	65.84	66.50
* WN	65.73	66.41

4 相关工作

事件检测一直是自然语言处理中的难点问题。传统特征驱动的事件检测方法通过一系列语法、语义相关(解析树、序列信息)的特征^[3]，来对事件检测进行建模。由于文本多样性和事件结构的复杂性，使得该类方法面临着繁重的特征工程和自然语言处理工具的错误传播问题。近年来，越来越多的研究者利用神经网络模型进行事件检测并取得了较大进展。CNN^[1,13]和双向 LSTM^[20-21]是目前两个通

① <https://catalog.ldc.upenn.edu/ldc2006t06>

用基础模型。不少工作在此基础上,利用事件检测和事件元素抽取进行联合建模获得更多的事件相关信息^[2],或者通过更加复杂的网络结构来建模更多的上下文信息^[22]。

由于神经网络模型的训练需要大量的标注语料,往往会有语料瓶颈的问题。为了得到更多的训练数据,近年来不少研究者利用远距离监督的方式,通过知识库直接生成事件弱标注数据。Liu 利用已有 ACE 数据分类器来对 FrameNet 的语义框架进行事件类别预测,并通过全局推理的方式来将语义框架与事件类别进行对应,以此扩展 ACE 的训练数据^[5]。Chen 则通过对 FreeBase 数据的核心元素进行检测、对触发词进行过滤和扩展,最终利用核心元素和触发词对 FreeBase 进行回标来生成召回率较高的自动标记数据^[4]。与上述工作相比,本文方法不需要引入新的文本,而是通过语法句子改写、语义词汇替换的方式来对原有训练数据进行扰动,来生成新的伪数据。

近年来,利用对抗样本来提升神经网络训练的鲁棒性也受到了越来越多的关注^[7]。但是,不同于图像数据输入的连续可导性,文本数据是离散不可导的,随意对词的替换可能会带来语义漂移或者产生不符合语法的文本。Li 等提出利用语言学的方式来对训练数据加入噪声,使得神经网络训练更加鲁棒^[9]。受该工作的启发,本文方法在尽可能保留事件语义的前提下,通过基于语法的句子改写和基于语义的词汇替换等方式生成相应伪数据来帮助训练。

5 结论与展望

针对深度学习模型在事件检测任务上的标注语料瓶颈和训练过程不稳定问题,本文提出了基于语言学扰动的伪数据生成方法。本文分别从语法和语义角度出发,提出了基于文本复述和句子压缩的句子改写方法,以及基于 WordNet 和词嵌入的词汇替换方法来生成伪数据。为了有效地利用伪数据,减少其给训练过程带来的噪声影响,我们提出了数据增强和多实例联合训练两种不同的训练策略。实验结果表明,基于语义的语言学扰动方法对模型的提升更加明显,同时基于多实例联合目标函数可以利用伪数据进行更加鲁棒的训练。

在未来的工作中,我们计划将基于语言学的扰动方法应用在其他的信息抽取任务(实体抽取、关系

抽取)中,同时借助先进的深度生成式模型来生成训练数据。

参考文献

- [1] Chen Y, Xu L, Liu K, et al. Event extraction via dynamic multipooling convolutional neural networks [C]//Proceedings of ACL 2015. Beijing, China: ACL, 2015: 167-176.
- [2] Nguyen H, Cho K, Grishman R. Joint event extraction via recurrent neural networks [C]//Proceedings of NAACL-HLT 2016. San Diego, USA: ACL, 2016: 300-309.
- [3] Li Q, Ji H, Huang L. Joint event extraction via structured prediction with global features [C]//Proceedings of ACL 2013. Sofia, Bulgaria: ACL, 2013: 73-82.
- [4] Chen Y, Liu S, Zhang X, et al. Automatically labeled data generation for large scale event extraction [C]//Proceedings of ACL 2017. Vancouver, Canada: ACL, 2017: 409-419.
- [5] Liu S, Chen Y, He S, et al. Leveraging framenet to improve automatic event detection [C]//Proceedings of ACL 2016. Berlin, Germany: ACL, 2016: 2134-2143.
- [6] Jiang Y, Zur M, Pesce L, et al. A study of the effect of noise injection on the training of artificial neural networks [C]//Proceedings of IJCNN 2013. Atlanta, USA: IEEE, 2009: 1428-1432.
- [7] Goodfellow I, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples [C]//Proceedings of ICLR 2015. Vancouver, Canada: 2016.
- [8] Zhao Z, Dua D, Singh S. Generating natural adversarial examples [C]//Proceedings of ICLR 2018. Vancouver, Canada: 2018.
- [9] Li Y, Cohn T, Baldwin T. Robust training under linguistic adversity [C]//Proceedings of EACL 2017 Valencia, Spain: ACL, 2017: 21-27.
- [10] Copestake A, Flickinger D. An open source grammar development environment and broad-coverage english grammar using hpsg [C]//Proceedings of LREC 2000. Athens, Greece: 2000: 167-176.
- [11] Copestake A, Flickinger D, Pollard C, et al. Minimal recursion semantics: An introduction [J]. Research on Language and Computation 2015, 3(2), 281-332.
- [12] Setiawan H, Dyer C, Resnik P. Discriminative word alignment with a function word reordering model [C]//Proceedings of EMNLP 2010. Massachusetts, USA: ACL, 2010: 534-544.
- [13] Nguyen H, Grishman R. Event detection and domain

- adaptation with convolutional neural networks [C]//Proceedings of ACL 2015. Beijing, China; ACL, 2015: 365-371.
- [14] Mrkšić N, Séaghdha DÓ, Thomson B, et al. Counter-fitting word vectors to linguistic constraints [C]//Proceedings of NAACL-HLT 2016. San Diego, USA; ACL, 2016: 142-148.
- [15] Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations [C]//Proceedings of NAACL 2013. Atlanta, USA; ACL, 2013: 746-751.
- [16] Heafield K, Pouzyrevsky I, Clark H, et al. Scalable modified kneserney language model estimation [C]//Proceedings of ACL 2013. Sofia, Bulgaria; ACL, 2013: 690-696.
- [17] Zeng D, Liu K, Chen Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks [C]//Proceedings of EMNLP 2015. Lisbon, Portugal; ACL, 2015: 1753-1762.
- [18] Zeiler D. Adadelta: An adaptive learning rate method [J]. arXiv preprint arXiv: 1212.5701, 2012.
- [19] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014, 15, 1929-1958.
- [20] Lin H, Lu Y, Han X, et al. Adaptive scaling for sparse detection in information extraction [C]//Proceedings of ACL 2018. Melbourne, Australia; ACL, 2018: 1033-1043.
- [21] Yang B, Mitchell T. Leveraging knowledge bases in lstms for improving machine reading [C]//Proceedings of ACL 2017. Vancouver, Canada; ACL, 2017: 1436-1446.
- [22] Lin H, Lu Y, Han X, et al. Nugget proposal networks for chinese event detection [C]//Proceedings of ACL 2018. Melbourne, Australia; ACL, 2018: 1565-1574.



陆垚杰(1993—),通信作者,博士研究生,主要研究领域为信息抽取及自然语言处理。

E-mail: yaojie2017@iscas.ac.cn



韩先培(1984—),博士,研究员,主要研究领域为信息抽取、知识库构建及自然语言处理。

E-mail: xianpei@iscas.ac.cn



林鸿宇(1993—),博士研究生,主要研究领域为信息抽取及自然语言处理。

E-mail: hongyu2016@iscas.ac.cn